Introduction to Text Mining in Big Data Analytics

Derrick L. Cogburn
American University
dcogburn@american.edu

Michael J. Hine
Carleton University
Mike.Hine@carleton.ca

Abstract

As a contribution to the HICSS 50th Anniversary Conference, we proposed a new mini-track on Text Mining in Big Data Analytics. This mini-track builds on the successful HICSS Workshop on Text Mining and recognizes the growing importance of unstructured text as a data source for descriptive and predictive analytics in research on collaboration systems and technologies. In this initial iteration of the mini-track, we have accepted three papers that cover conceptual issues, methodological approaches to social media, and the development of categorization models and dictionaries useful in a corporate context. The mini-track highlights the potential of an interdisciplinary research community within the HICSS collaboration systems and technologies track.

1. Introduction

The Text Mining in Big Data Analytics Mini-Track is a new addition to the Collaboration Systems and Technologies Track at the Hawaii International Conference on System Sciences (HICSS). It was developed as a contribution to the 50th Anniversary HICSS Conference to build on the successful two-time HICSS workshop on text mining, in recognition of the growing focus on big data analytics within HICSS, and the reality that global collaboration systems, social media, and information systems of all types, generate enormous amounts of unstructured textual data. These types of data include system logs, email archives, websites, blog posts, meeting transcripts, speeches, annual reports, published material, and social media posts. While this unstructured textual data is readily available, it presents a tremendous challenge to researchers trying to analyze these large bodies of text with traditional social science methods.

Text mining in big data analytics is becoming increasingly important to an interdisciplinary group of scholars, practitioners, government officials, and international organizations. For example, the American Association for the Advancement of Science (AAAS) launched a new competition in 2014 on Big Data and Analytics within its highly competitive senior executive branch fellowship program. Other corporate initiatives like the Big Boulder Initiative was formed in 2014 (http://bigboulderinitiative.org) as a trade association wholly dedicated to "the advancement of social data in businesses and organizations of all kinds".

However, within this growing focus on big data and text mining, there is a dilemma. While as much as 75-80% of available data is unstructured text, many social scientists are not trained in the techniques for analyzing large bodies of text, but there is a growing effort to reverse this situation and this mini-track is designed to accelerate this process and to integrate with the HICSS Workshop on Text Mining to build an interdisciplinary text mining research community, particularly focused on textual data emerging from collaboration systems and technologies.

The HICSS Text Mining mini-track invited papers that apply text-mining approaches to a wide variety of substantive and interdisciplinary domains, including, but not limited to theoretical and applied approaches to analyzing:

- Blog posts
- Twitter and social media analysis
- Email archives
- Published articles
- Websites and blogs
- Meeting transcripts
- Speeches

Addressing methodological challenges, such as:

- Automated acquisition and cleaning data
- Working on distributed, high-performance computers
- Overcoming API limitations
- Using LDA, LSA, and other techniques.

As co-chairs of the HICSS Text Mining Mini-Track, we are pleased with the results of this initial offering. We have accepted three papers that highlight various important aspects of this emerging community. The first paper by Chelsea Hicks addresses some of the key conceptual issues in text mining for collaboration

URI: http://hdl.handle.net/10125/41262
ISBN: 978-0-9981331-0-2

systems and is entitled, "An Ontological Approach to Misinformation: Quickly Finding Relevant Information". Next, recognizing the importance of social media to collaboration systems and technologies, our second paper by Yuheng Hu and Yili Hong focuses on the potential of social media for predictive analytics and is entitled, "Modeling Twitter Engagement in Real-World Events". In a focus on the potential of text mining in analyzing and understanding corporate environments, our final paper, by Qi Deng, Mike Hine, Shaobo Ji, and Sujit Sur, is entitled, "Building an Environmental Sustainability Dictionary for the IT Industry". Each of these papers will be summarized below, followed by implications of these papers for the potential of an interdisciplinary research community within HICSS on text mining and big data analytics.

2. Approaches to Misinformation

The first paper in our mini-track is by Chelsea Hicks. This paper addresses some of the key conceptual issues in text mining for collaboration systems and is entitled, "An Ontological Approach to Misinformation: Quickly Finding Relevant Information". This paper sees identifying misinformation or rumors as an important and growing field of research in information systems. It cites several examples of widespread misinformation spreading during national crises, such as the Boston Bombings and during the Ebola crisis in West Africa. During these crises, rumors spread quickly on a wide variety of social media platforms. These rumors can obfuscate the reality of these situations and accelerate the spread of even more rumors. This study uses the semantic web to address this problem, and proposes that ontologies can help find accurate information for a case quickly and accurately. The paper develops a weighting formula to help identify and display the most relevant results to an interested party, and outlines plans for implementing this approach on appropriate datasets.

3. Modeling Twitter in Real World Events

The second paper in our mini-track recognizes the importance of social media to collaboration systems and technologies. It is written by Yuheng Hu and Yili Hong, and focuses on the potential of social media for predictive analytics and is entitled, "Modeling Twitter Engagement in Real-World Events". In this paper, the authors argue that people invest substantial time, attention, and emotion while engaging in various activities in the real world. They see social media platforms such as Twitter, offering tremendous opportunities for people to become engaged in real-world events (e.g., a product release conference, news conference, and political elections, through information sharing and communicating about these events. However, the authors argue that there is insufficient understanding of the factors that affect people's Twitter engagement in such real-world events. This paper addresses these issues by first operationalizing a person's Twitter engagement in real world events. In this operationalization, it includes individual activities such as posting, retweeting, or replying to tweets about such events. The authors then construct statistical models that examine multiple predictive factors associated with four different perspectives of users' Twitter engagement with 643 real-world events. The authors find that the measures of people's prior Twitter activities are all variously correlated to their engagement with real-world events.

4. Text Mining Dictionaries for Industry

In a focus on the potential of text mining in analyzing and understanding corporate environments, our final paper, by Qi Deng, Mike Hine, Shaobo Ji, and Sujit Sur, is entitled, "Building an Environmental Sustainability Dictionary for the IT Industry". This paper builds on the content analysis approach, already used commonly within corporate sustainability research. However, in line with the broader argument made in this mini-track, that much more unstructured textual data is available for analysis than is currently being utilized by most researchers who rely only on human coding, this study seeks to overcome those limitations by focusing on text mining techniques. Specifically, this paper develops a domain specific dictionary designed to analyze corporate sustainability reports for the IT industry. The paper also presents a standardized dictionary building process model, applicable to multiple domains.

5. Towards a Text Mining Community

We believe this new mini-track has great potential to stimulate the creation of a robust, interdisciplinary text mining research community within HICSS. Given the amount of unstructured textual data generated by widespread collaboration systems and technologies, such a research community would be invaluable. The text mining papers at this 50th Anniversary HICSS Conference represent what we see as an important emergent trend, which we believe will remain for many years to come.